L.A.S.L.A. and Collatinus: a convergence in lexica

Philippe Verkerk, Yves Ouvrard, Margherita Fantoli, Dominique Longrée

L.A.S.L.A. and Collatinus: a convergence in lexica

Philippe Verkerk, Yves Ouvrard,
Margherita Fantoli and Dominique Longrée

L.A.S.L.A. (Laboratoire d'Analyse Statistique des Langues Anciennes, University of Liège, Belgium) has begun in 1961 a project of lemmatisation and morphosyntactic tagging of Latin texts. This project is still running with new texts lemmatised each year. The resulting files have been recently opened to the interested scholars and they now count approximatively 2.500.000 words, the lemmatisation of which has been checked by a philologist. In the early 2.000's, Collatinus has been developed by Yves Ouvrard for teaching. Its goal was to generate a complete lexical aid, with a short translation and the morphological analyses of the forms, for any text that can be given to the students. Although these two projects look very different, they met a few years ago in the conception of a new tool to speed up the lemmatisation process of Latin texts at L.A.S.L.A.. This tool is based on a concurrent lemmatisation of each word by looking for the form in those already analysed in the L.A.S.L.A. files and by Collatinus. This lemmatisation is followed by a disambiguation process with a second-order hidden Markov model and the result is presented in a text-editor to be corrected by the philologist.

1 L.A.S.L.A.

The Laboratory for Statistic Analysis of Classical Languages (L.A.S.L.A. in the following) was founded in November 1961 at the University of Liège, by L. Delatte and E. Évrard. Its original aim is to lemmatize and analyze (tag) literary classical texts, both in Greek and in Latin, in order to produce indexes and to allow the study of classical languages with statistical and quantitative methods. This project is still going on with, as an outcome, large digitalized, lemmatized and annotated Latin corpora. These corpora cover the classical period, from Plautus to Ausonius, with some other Late-Latin texts. The L.A.S.L.A. Encoding Initiative interface allows to add new texts to the corpora. The L.A.S.L.A. also released Textual Data Analysis tools to access the information contained in its files. Through a specific agreement, the access to these files is now free and open for every scholar who asks for it.

1.1 The structure of the files

The L.A.S.L.A. Latin files contain the fully lemmatized texts with a complete morphosyntactic analysis and some syntactic information. They have been systematically verified by a confirmed Latinist (either M.A. or Ph.D.). The annotation is not related to any specific grammar or to any specific linguistic description. In short, the available files are put in a text format where each line contains all the information related to a single token. As a reminiscence of the old punched cards, the fields have a fixed length, the blank character filling the empty spaces.

Each line begins with an alphanumeric code that refers to the considered text and a number that counts the sentences. Any punctuation, added anyhow by the modern editors, is removed, except for the period that separates the sentences. For each token of the text, the line contains the

lemma –as it appears in the dictionary of reference[1]– associated with an index if there are different homographs or to spot proper names or their derived adjectives. Then come the form as it appears in the text, the reference –according to the *ars citandi*– and the complete morphologic analysis in an alphanumeric format[2]. For the verbs, an extra field (which remains empty for any other Part-of-speech, shorten in PoS in the following) gives some syntactic information: the verb of the main clause is singled out and a subordinate code –depending on the subordination type– is affected to the other verbs in the sentence.

The lemma always refers to an entry in the Forcellini's dictionary with a systematic disambiguation. For instance, POPVLVS_1 is the people, while POPVLVS_2 is the poplar. The PoS is also used to distinguish the homographs as AMICVS_1, the substantive, and AMICVS_2, the adjective. A problem arose for late Latin texts where an adjective can become a substantive. This is the case for SANCTVS, which is only an adjective in classical Latin, but became a substantive later, especially in religious texts. To handle this situation an extra tag has been introduced "use as a substantive".

During the tokenisation process, the enclitics are separated from the rest of the form, but a special character is inserted in the line to remind that those two tokens give a single word. At the opposite, the encoding allows to treat verbal compound forms and also the ellipsis. The crasis is treated in a way quite similar to the enclitics: one word leads to two lemmata. The tmesis and the compound words are also encoded in a special way.

The morphologic tag in 9 characters begins with the PoS in one letter (A=noun, B=verb, C=adj. etc...) which is followed by a figure indicating the declension (for a noun), the group (for a verb) or the class (for the adjectives). Then come as one digit for each, if relevant, the case, the number, the degree, the mood, the tense, the voice and the person. For the same lemma, the figure indicating the declension can vary. For instance, *Vlixes* belongs, in principle, to the third declension. However, in accusative singular, the two forms *Vlixem* and *Vlixen* exist and are associated to different tags: A331 for the first one, as it is the normal Latin form, and A731 for the second form which is the Greek one. For the genitive, the two forms *Vlixi* and *Vlixei* are characteristic for the second declension, so the tag is now A241, although the lemma is still VLIXES. The gender is an extra piece of information but, due to the original choice made by the founders, it is not given for the nouns and is not fully disambiguated. As a matter of fact, there are six possible genders according to the L.A.S.L.A. files[3].

1.2 The L.A.S.L.A. Encoding Initiative interface

The L.A.S.L.A. Encoding Initiative interface is mainly a selection interface. A new text is first given to an operator who proceeds to some preprocessing[4]: tokenization, lemmatization and analysis. Until 2019, this was achieved with an analytic lemmatizer : the form were decomposed from its ending in order to get all the possible roots and then recomposed in order to get all the possible analysis for all the possible lemmata. However the software supporting this lemmatizer

1 The *Lexicon totius latinitatis* of Forcellini, edited by Corradini, Padua, 1864.

2 As a matter of fact, two alphanumeric encodings co-exist, one in 5 characters –which is the original one– and the other with 9 –which is simpler. The matching can be done automatically.

3 The three real genders and the three combinasions that exist in the declensions (the f. + n. combinasion does not exist). We have in project (not yet fully engaged) to add the gender of the nouns and to desambiguise, when possible, the gender of the adjectives, depending on the associated noun. Part of the task can be done automatically, but the result will have to be checked. Some words, as *canis* or *pereger*, are common (both masculine and feminine) and, even with the context, it may happen that the gender cannot be decided for sure.

4 cf. Denooz 1978 and Philippart de Foy 2014

was obsolete and L.A.S.L.A decided to build a new lemmatizer based on form recognition. It means that each word of the text is compared to all the forms present in the L.A.S.L.A. forms dictionary. This forms dictionary includes all the possible forms for all the lemmata included in the L.A.S.L.A. lemmata dictionary. These possible forms are generated with a software based on morphologic rules, which adds all possible endings to each root corresponding to a lemma from the L.A.S.L.A. lemmata dictionary. The limitation here is that only lemmata already met in treated texts are included in the dictionary.

After this preprocessing, the text is presented as a list of tokens followed by all the known lemmatizations and analyses, one per line, in the alphanumerical order of the tags (corresponding to a given and fixed order of PoS, PoS-subcategories and morphosyntactic categories). Then, the philologist comes into play by selecting the "good one". He/she is also invited to enter the syntactic information for the verbs, as it cannot be guessed by the computer. If a form is not in the dictionary or if the proper analysis is not given, the philologist has to add the right analysis. The validation of the annotated text is possible only when the philologist has selected one analysis for each form of the text. At the end, the treated text returns to an operator who puts it in its final form.

Such a procedure ensures that the philologist has checked the lemmatisation and the analysis of each token. As the computer does not select a priori a solution (even if there is only one possible lemmatisation and analysis), the philologist has to read really every line on the screen. However, it has also its drawback, especially for technical texts using a specific vocabulary. As the dictionary has been built on "literary" texts, a large amount of scientific words are missing and the philologist has to add them one by one. Moreover, the text being prepared a priori with the data already known, there is no way to copy automatically the new analysis to the same form which could appear further in the text. As a matter of fact, to guarantee the coherence of its dictionary, the L.A.S.L.A. does not update it automatically with the new forms.

1.3 Access to the information

There are several ways to access the information stored in the L.A.S.L.A. files. The simplest approach is given by the interface "Opera Latina"[5] which allows for documentary search (indexes) but gives no statistics. A second possibility is to download the package "Hyperbase-Latin" which allows documentary and statistical exploration. This software has been developed in collaboration with Étienne Brunet of the laboratory "Bases, Corpus, Langage" (UMR 7032; CNRS-University of Nice).

A more flexible approach is offered by the interface Hyperbase Web Edition[6]. One can choose between various databases or corpora. Beyond the usual documentary search (indexes), one can also ask for pattern detection, for instance all the sequences of two nouns. Hyperbase Web Edition allows statistic search as z-score, factorial analysis or tree analysis. It is also possible to study the co-occurrences and even co-occurrences of pairs. As an extension of Hyperbase Web Edition, HyperDeep, which is based on a Convolutional Neural Network, allows to identify what is characteristic of a text or to find influences between authors.

For more specific purposes, the L.A.S.L.A. files can be converted to XML and treated with TXM[7] or with data-mining tools[8].

5 The list of the available texts is given at http://web.philo.ulg.ac.be/lasla/textes-latins-traites/

6 http://hyperbase.unice.fr/hyperbase/?edition=lasla

7 textometrie.ens-lyon.fr/spip.php?rubrique96

8 https://tal.lipn.univ-paris13.fr/sdmc/

2 Collatinus

Collatinus[9] has been originally developed by Yves Ouvrard for teaching. It allows to generate a complete lexical aid, with a short translation and the morphological analyses of the forms, for any text which can be given to the students. With time, the lemmatizer has been complemented with other useful tools[10]. By simply clicking on a word, one can open a digital dictionary, as Lewis & Short or Gaffiot, to have the complete definition of the lemma. Another possibility is to scan a text to identify its metrical structure. A probabilistic tagger, based on a second order hidden Markov model (shorten as HMM in the following), allows to choose the best lemmatization and analysis for each form taking into account the context.

The lemmatization of a form is obtained by trying to split it as a root associated with a standard word-ending, which reproduces what the human reader does. The advantage of a program as Collatinus is that it is able to recognize forms not yet seen as soon as the root-word is known[11]. It is also easier to improve its base of knowledge: adding the data for a new root-word allows to recognize immediately ten or more (even a hundred, for verbs) forms. Obviously, a program as Collatinus "knows" a lot of forms that are not attested in the texts that have survived[12].

2.1 Principle of operation

When a student learns Latin, the first thing he/she has to understand is the way forms are constructed. Words are connected to an inflection paradigm. For each paradigm, one has to learn the list of word-endings and the rules to combine these endings with the roots that can be calculated, in some cases, or must be given. Collatinus works exactly in this way: it has one file that gives the word-endings and the construction rules for each paradigm and another one that connects the lemmata to the paradigms and gives also the roots which cannot be calculated. With this data, the construction of the inflected forms is immediate.

However, the lemmatization of a form requires the reverse process. For a given form, we have to split it in all the possible ways and to check that the first part coincides with a known root and the last one with a word-ending associated to the paradigm of the root[13]. The word-endings carry part of the information for the analysis, which is then stored in the file. Instead of an explicit analysis as e.g. *nominative singular*, we just made a list of the morphosyntactical analyses, which are possible in Latin and coded the analysis with a simple number. As a matter of fact, these analyses are only 416. The number is converted in its human readable form when needed, i.e. for the display. By the way, this encoding allows also translating the analysis in different languages[14].

---

9 "*Collatinus, un outil polymorphe pour l'étude du latin*" by Y. Ouvrard and Ph. Verkerk, in Archivum Latinitatis Medii Aevi, **72**, 305-311 (2014)

10 For more details about these functionalities, see the article on Collatinus in the Proceedings of Digital Text Analysis III, Heidelberg 2017 (to be published, if ever, on Classics@, available as preprint https://hal.archives-ouvertes.fr/hal-02385036).

11 For any unknown form coming from an unknown root-word, it should be possible to guess a reasonable root-word in some simple cases.

12 Note that, if the classical corpus is well established, it is not the case for medieval Latin.

13 Going further, one can imagine to guess the lemma simply by substracting the common word-endings. However, it would lead to surprizing results. For instance, the form *merobibus* could be analyzed as an ablative plural of an hypothetical *merobis*. But such a method could give good results if several forms of the same lemma are found in a text.

14 For the moment, French, English and Spanish. But one can convert it to any other computer-oriented forms.

2.1.1 First difficulties

One of the aim of Collatinus is to treat a Latin text as it is, without requiring some preprocessing steps as the tokenisation. A difficulty appears because of the enclitics *-que*, *-ne* and *-ve*. These words are glued at the end of any form, and have to be separated for the lemmatization. In most of the cases, the enclitics *-que* and *-ve* do not lead to ambiguous forms[15], which is not the case of the enclitic *-ne*. For instance, a form as *mentione* could be analyzed as the ablative singular of *mentio, onis*, as well as the nominative followed by the enclitic *-ne*. However, the enclitics are not so frequent, thus we assume that if a form can be lemmatized as it is, then it is not necessary to search for the enclitics. In other words, the form *mentione* is now analyzed only as the ablative of *mentio*.

Collatinus also knows some contraction and assimilation rules. For instance, it is frequent that a double "i" appearing in the flexion of a word[16] is written as a single long-i. Some forms of the perfectum can be contracted, the *-vi-* disappearing in, for instance, *amasse* (for *amavisse*). These forms are recognized by Collatinus, without the necessity of adding new word-endings. For the verbs constructed with a prefix, assimilation can change the spelling in some cases. It is the case, for instance, of *adfero, adtuli, adlatum* which often becomes *affero, attuli, allatum*[17]. The main assimilations of the prefix are known by Collatinus and built-in, so that it avoids the proliferation of forms for the same word.

2.1.2 Distinction between *u* and *v*

Very often, the Latinists do not distinguish the letters u and v, and erase the j from the alphabet. However, for counting the syllables or the meters, it is clearly necessary to make the distinction. Thus, Collatinus keeps, in its lexicon and in the word-endings, the two consonants "v" and "j", said to be Ramist consonants[18]. By the way, if one wants to use only "u" and "i", it is easy to replace "v" by "u" and "j" by "i". The proof, if needed, that the distinction is the best choice is that the reverse process (restoring "v" and "j") is almost impossible, at least very difficult, except through a lemmatization method.

On the other hand, several Latin texts use only the "u" and "i", and Collatinus knows it[19]. The way to solve the problem is obtained with two steps. In a first step, all the "v" are replaced by "u" for the lemmatization. Then in a second step, the form is reconstructed from the root and the word-endings that eventually contain the "v" and "j". As a result, a word as *uoluit* is analyzed as a form of perfect of either *volo* or *volvo*[20]. However, if the text contains *voluit*, with a "v", one can assume that it is not the perfect of *volvo*, otherwise it should have been written *volvit*, with two "v".

---

15 A noticeable exception is "quo-que" that appears 7 times in the texts lemmatized by the L.A.S.L.A. (to be compared to the 2.290 occurrences of the lemma "quoque").

16 The first i ending the root, often short, and the second one at the beginning of the word-ending combine in a long-i.

17 Gaffiot gives the first forms, while Lewis and Short prefers the second ones.

18 Pierre de la Ramée (Petrus Ramus) is known in France to have introduced this distinction u/v and i/j in his "Gramere" (1562). But it seems that this idea appeared earlier in Spain (Antonio Nebrija, 1492) or in Italy (Giovanni Trisino, 1529). See Xavier Blanco i Escoda et Krzysztof Bogacki, *Introduction à l'histoire de la langue française*, Bellaterra, Université autonome de Barcelone, coll. « Documents » (n 104),2014, p. 160, n. 24 and p. 161.

19 In the worst case, the editors write the capital-U as V. It is not unfrequent to find Vnde at the beginning of a sentence or to meet Vlixes in some texts.

20 *volvit* can also be a form of the present of *volvo*. The meaning of the sentence allows the reader to identify the good form, but a computer does not understand the text. The case of *uoluit* can be a problem in prosody as it can counts for two or three syllables.

If the form of the text contains one (or more) “v”, the program eliminates any lemmatisation that would lead to a reconstructed form with a different number of “v”.

An other group of “u” are not “real” vowels: it is the case of *su̯avis* and of *sangu̯is*. It is also the case for the group “qu”, but in this last case the “u” is never a vowel. In the groups “sua” or “gui”, there are examples where the “u” is a vowel, for instance the possessive *sŭă* and the adjective *āmbĭgŭīs*[21]. It would have been somehow shocking to write *svavis* or *sangvis* to stress that these words have only two syllables. Instead, we use the expunctuated-u and write *sụāvĭs* and *sāngụĭs*[22].

2.1.3 Word-endings and construction rules

As already said, besides the lexicon which will be discussed later, Collatinus has an other important file which gives the word-endings and the construction rules. For each paradigm, it gives the list of analyses and the corresponding word-ending. A noun that follows a usual declension has 12 analyses and word-endings (some of them are identical), while an adjective has 108 possible analyses and word-endings. All the possible combinations of case, number, gender, degree, tense, mood and voice give 416 analyses which are just designated with a number. To avoid a very long enumeration of word-endings, we introduced a mechanism by which a paradigm “inherits” the endings of its parent[23]. For instance, *miles* and *civis* have most of their endings in common, so we just have to indicate the differences.

Obviously, the word-ending is not the end of the story because one has to know the root to which this ending can be appended. For some declensions or conjugations, the roots can be calculated with the sole lemma. For instance, for the first declension, it is sufficient to drop the last character of the lemma to have the root. In other cases, it must be given by the lexicon: one cannot guess the root *mīlĭt-* for the lemma *mīlĕs*. A more subtle example is the case of the first conjugation. In most cases, the roots for the perfect and the supine are obtained by adding “āv” and “āt” to the main root: the knowledge of the form *ămo* is sufficient to calculate the three roots *ăm, ămāv and ămāt*, so it is not necessary to give them in the lexicon. But some verbs of the first conjugation do not follow this simple construction rule. To solve this problem, we have decided that if a root is given in the lexicon, it replaces the one that could be calculated. For instance, for the verb *sŏno*, we give the two roots *sŏnŭ-* and *sŏnĭt-* for the perfect and the supine.

2.1.4 Ordering of the solutions

For several forms, the result of the lemmatization is not unique[24]. Different words can lead to the same form, or a form corresponds to different analyses of the same word. It may be interesting that Collatinus gives the different solutions in an order that reflects the frequency of the use of the words. Up to version 10, the order of the solutions was alphabetical. As a result, the lemmatization of *suis*, for instance, gave the genitive of *sus, suis* as the first solution, although the ablative or the dative of *suus, a, um* are more likely.

---

21 The vowels are marked with a macron “¯” when they are long, as *ā* or *ī*, and with a breve “˘” when they are short, as *ĭ* or *ŭ*.

22 Once again, if one does not want to use this strange character, it is easy to replace it by the standard “u”.

23 The construction rules are also transferred.

24 There is a problem of vocabulary around the lemmatisation: for the final user, the aim of a lemmatizer is to give **the** (unique) lemma associated to a given form in a given sentence. However, the operation consisting in giving **all** the lemmata that can be associated with a form is also a lemmatisation. We prefer to stick to this last sense and the full process with the association of a single lemma to a form is obtained with two steps: lemmatisation and desambiguisation.

Thanks to the statistics made on the lemmatized texts of the L.A.S.L.A., we are now able to associate to each word of the lexicon a number of occurrences. Obviously, this number of occurrences is limited to the lemmatized corpus, but one can consider it as representative for the frequency of the words. To go back to the previous example, *sus* appears 47 times in the texts of the L.A.S.L.A., while *suus* appears 7,120 times. As Collatinus is not a form-lemmatizer[25], it does not know the number of occurrences for *suis* as dative plural of *suus* and for *suis* as ablative plural of the same *suus*. To order these two possible solutions, we make a strong assumption: the usage of the cases and number[26] (for nouns and adjectives; replaced by the mood for verbs) does not depend on the particular word. We still take into account the PoS[27] of the word. This evaluation does not reproduce exactly the observed frequencies, but remains a fair approximation. There are noticeable exceptions: for instance, *patres* is mainly a vocative plural, a case that is only very seldom used in other nouns/adjectives.

This ordering of the solutions is not sensitive to the context. Its depends only on the form itself and its analyses. According to the statistics done on the lemmatized text of the L.A.S.L.A., choosing the most frequent analysis gives the good result in 80% of the cases. To reach a lower error rate, one can develop disambiguation methods based on the tagging of the words. These methods take into account, very crudely, the context of the word. They will be discussed later.

2.2 Extension of the lexicon

The lexicon of Collatinus contains the lemmata associated to a known paradigm, the different root-words that cannot be calculated and various pieces of information, as the number of occurrences of this lemma in the texts lemmatized by the L.A.S.L.A.. The translations of these lemmata are given in distinct files (one for each language) so that the material necessary to inflect or analyse the forms is independent from the translations. It also allows to add more languages for the translations without having to duplicate or to change the basic information that rules the inflection. The files are just plain text-files, so that they can be edited and modified by the user to give better results.

Up to its version 10.2, the lexicon of Collatinus was set-up manually, the words being typed in when they were found in new texts given to the students. It was containing slightly less than 11,000 entries, which allow to lemmatize a significant part of the classical texts. However, we have decided to boost it by working on the dictionaries in a digital form. The two main dictionaries we have used are Lewis and Short (L&S), converted in XML by the Perseus Project[28], and Gaffiot, converted in TeX by a team lead by Gérard Gréco[29]. We have also used Georges[30] and Jeanneau[31] in their HTML forms. All these dictionaries are part of Collatinus. Some extra pieces of information were also used[32].

---

25 We shall come back later on that example through the L.A.S.L.A. tagger.

26 Unfortunately, the lemmatization by the L.A.S.L.A. does not give precisely the gender of the adjectives.

27 Mainly: noun, adjective, verb and pronoun, as categorized by the L.A.S.L.A..

28 Charlton T. Lewis & Charles Short, *A Latin dictionary founded on Andrew's edition of Freund's Latin dictionary*, Oxford 1879, encoded in XML by Perseus http://www.perseus.tufts.edu/

29 http://gerardgreco.free.fr/spip.php?article47 *Dictionnaire Latin-Français de Félix Gaffiot* (1934) by Gérard Gréco, Mark De Wilde, Bernard Maréchal and Katsuhiko Ôkubo. Thanks to Gérard Gréco, we had access to the file before its publication.

30 Karl Ernst Georges, *Ausführliches lateinisch-deutsches Handwörterbuch*, Hannover 1913.

31 Gérard Jeanneau, http://www.prima-elementa.fr/Dico. This Latin-French dictionary is still evolving. For this work, we have used a version of 2013.

32 The data from Collatinus itself, a short version of Gaffiot, An Elementary Latin Dictionary (English) by Charlton T. Lewis, and the headwords of the Pocket Oxford Latin dictionary.

The first part of the work has been to collect all the lemmata together with the morphological information and the translation in each dictionary. The precise tagging of L&S and of Gaffiot, although very different, allows to compile very rich databases. The translations were probably the most difficult part of the job. Subentries, as adjectives that derive from a noun which is the headword, were collected too. The graphical variants, often indicated in an abbreviated form as, for instance, *affĕro (better adf-)*, were expanded and added to the base. This has been done programmatically but checked afterwards. The internal variants, for instance *rĕverto (-vort-)*, have been especially difficult to treat, although they are rather intuitive for the human reader. Obviously, one has to face the imperfection of the tagging[33]: some tags are missing or do not include all the relevant information.

To deal with this lack of information, we combine the databases coming from the various dictionaries, with the idea that if a supine-form is missing in L&S, we can find it in Gaffiot (or vice-versa). This combination requires a kind of alignment of the files, especially for the homonyms, and the elimination of the redundant doublets. For instance, in L&S, *abscisus* has its own entry with a laconic definition "P. a., v. abscido" and is translated in a subentry of *abscido*. A supervised program allowed us to do that in a reasonable amount of time. The quantities can be sufficient to distinguish the homonyms as *pŏpŭlus* vs *pōpŭlus*, but not always. Sometimes, we have to consider the Part-of-Speech, as for instance in *a-spergo, ersi, ersum, 3, v. a.* vs *aspergo, ĭnis, f.*, or the gender to recognize the homonyms, for instance the noun *par, paris* which can be masculine or neuter. As a last chance, the human reader can use the translations to align the entries.

The last step is to convert the collected information in a file which can be understood by Collatinus. The quantities given by the dictionaries are compared, and sometimes they differ in which case we choose the form given by the "majority"[34]. The quantities that can be determined by position are usually not indicated, but the program knows the rules[35] so that it was able to supply the missing quantities to Collatinus. Once again, a difficult step is the reconstruction of the roots with the abbreviated indications. For the verb *a-spergo*, the program builds the form *āspērgo*[36] and the two roots, for the perfect and the supine, *āspērs*[37]. While for the noun, it gives *āspērgŏ*[38] and *āspērgĭn*.

This treatment, mostly automated, leads to a lexicon of about 77,000 lemmata, associated with a paradigm and the necessary roots. But some 7,200 more words have been extracted from the dictionaries and were not "understood". Some of them are useless for Collatinus: for instance, Gaffiot and the elementary Lewis have an entry for *aberam*, which is not a fundamental word. A latinist should go through this file to determine which words may be useful to complete the lexicon. On the other hand, the process of expanding the variants of the headwords, which was necessary to align the entries of the dictionaries[39], leads to doublets. Most of those due to the assimilation of a

[33] Here, we are considering the XML/HTML tags that identify the different entities. Later on, the word "tag" will have a rather different meaning.

[34] In the comparison of the quantities, we have to take into account that Georges and the Elementary Lewis indicate only the long vowels. The unmarked vowels can be either long by position or short.

[35] A diphthong is usually long (except for the æ of *præ* before a vowel, which becomes short). A vowel placed before two or more consonants is long too. A vowel before an other vowel is short.

[36] The quantity of the final o is not relevant, because it is given by the word-endings.

[37] In these case, the two roots are equal, but they usually differ. A difficult example is *ab-sorbĕo, bui, rarely psi, ptum* where we have two different roots for the perfectum, ābsōrbŭ and ābsōrps.

[38] The rule that says that the final o of the nominative is long when the previous vowel is long (L. Quicherat, *Nouvelle prosodie latine*, 30e édition, Paris 1885, p. 32, which can be downloaded from Gallica) seems not well followed. We prefer to mark it as common.

[39] For instance, Gaffiot has *adfero* as a headword, while L&S gives *affero* with the variant *adf-*. Both merge in Collatinus to give a single entry.

prefix have been tracked down and suppressed. Some Greek names are also Latinized (e.g. *Ariadna, ae* for *Ariadne, es*) leading also to doublets. But a similarity a/e or us/os is not sufficient: for instance, *Agylla, ae* is an Etrurian city, while *Agylle, es* is a nymph. A last group of doublets comes from the singular or plural forms of some words which are chosen as headwords in the different dictionaries. A careful hunting of all these duplicates is still to be done.

At the end, to avoid long loading times, we split the lexicon in two parts. About one third of it corresponds to the 24,000 words that have been found in the texts lemmatized by the L.A.S.L.A.. It is loaded by default and allows the lemmatization of a large fraction of the classical texts. The remaining two thirds, 53,000 words, are rarer words and are loaded only on demand. We had also the project to split the lexicon in more parts, each one specialized in a period of time or a range of semantically similar topics. We are considering this possibility for future versions as it requires that the program is able to load and purge different lexica while running[40].

2.3 Perspective – Modularity of the data.

The 12th version of Collatinus (C12 here) is still under development. It focuses essentially on lexical and morphological data. Its aim is to handle larger and more precise data to lemmatize specialized corpora.

For instance, having to lemmatize a large medieval corpus, we have been confronted to a couple of difficulties.

- Numerous new words
- Evolution of semantics
- Evolution of graphic uses
- Evolution of paradigms

So, we found that the actual state of Collatinus' data often leads to wrongs results.

2.3.1 Modules

The idea is to collect all the differences between the classical data and those which are required to lemmatize a non-classical corpus, for instance a medieval one. Using a special editor, a new set of data is created, containing all the differences between the classical state of the Latin language and the one in the corpus under study. These differences may appear at various levels: lexicon and translations, inflections, graphic usages, irregular forms. This data is zipped into a package with the *.col extension. Once created, this module can be uploaded to the web-site of Collatinus. Then, other users can download it and install it in their C12.

So, having to lemmatize a medieval text, the C12 user selects the medieval module. First of all, C12 reads classical data. Then, from this medieval module, new words are added. If a word already exists in classical data, it is replaced by the medieval one. Often, the medieval word has few differences with the classical one: for instance, just a new meaning. Sometimes, a word only needs to change its flexional paradigm, or one of its stems. But it may also be completely different. The same principle is applied for inflexions, irregular forms and graphic variants.

Graphic variants: C12 adds a new data file, named *vargraph.la* which stores the graphic particularities:

- Classical ones, e.g. cu/quu (cum/quum).
- Medieval graphic variants are numerous, e.g.:

[40] For the moment, Collatinus loads the data when booting.

. ligatures q;/que,
. phonetics mpn/mn (dampnum); ß/ss
. tilde: ã or ā/an, am

For medieval modules, the problem of the lexicon is very acute. Medieval corpora introduce many anthroponyms, toponyms, latinization of local words: Celtic, Germanic, Spanish etc. And these new words depends strongly on the considered corpus. For instance, the words derived from the vernacular languages will differ in Spain and in Germany. Thus, specific specialized lexica may be needed for each corpus[41].

A real difficulty is the survival of the anterior states of language. Classical authors could not know words to be created during the following centuries, but ulterior authors did know, sometimes very well, classical authors. We need to be very careful when editing a classical word: classical senses may survive in medieval texts.

2.3.2 The editor: Ecce

Ecce (Ecce Collatinistarum Communitatis Editor) aims to create modules for C12. Ecce's interface has four tabs: Lexical Modules, Lexicon, Graphic variants, Irregulars. When launched, the first tab, Lexical Modules, is selected. On the left part, the user can choose the module to activate, deactivate, delete, generate or install. He can also choose other modules to dig data he will be able to add to the new module. Let us call them 'tank modules'. A very important tank is lem_ext, named 'extension'. When the module to feed and tank modules are selected, the user clicks the 'Activate' button. If this modular approach is adopted and widely used, the number of tank modules will grow, and building new modules will be easier and easier.

The Lexicon tab appears. Latin text, and navigation buttons: beginning, backward, forward, previous failure, next failure, end. To feed the lexicon, the user clicks the 'next failure' button. Ecce goes on lemmatizing the text word after word, and stops when the lemmatization fails. The word is displayed, solutions, if any, are searched in tank modules, so that you can check them, edit one of them, and add it. You can also, on the right part, edit a new lemma from scratch. If the lemma exists with another spelling, or another flexion, the two other tabs can be used. When the new data is validated, it is a good practice to go back to the beginning, and restart the lemmatization, to check if the edition is right.

2.3.3 Usages

Collatinus is a lemmatizer, and its main usage is lemmatization. The modular organization of C12 allows a more precise lemmatisation of non-classical or special corpora: author, place, topic. Like Mario Nizzoli, in 1734, released a *Thesaurus Ciceronianus*, a Ciceronian C12 module could be created, uploaded to the web site of Biblissima and then downloaded by any other user who may be interested.

It could be interesting to test it for teaching tasks:

- Provide a tiny module for a short latin text;
- Ask students not to translate a text, but to develop the module which fits to this text, using Ecce.

[41] Another possibility would be to use an expandable personal lexicon, but it would remain “private” and every scholar would have to develop their own lexicon. A third way could be to gather a huge data-base, but at some point a trade-off has to be found between the size of the base and the responsivity of the programme.

3 L.A.S.L.A.-Tagger

As in every language, forms in Latin can be ambiguous. This ambiguity can be at different levels. On one hand, in a declension, different cases can have the same form for the same word. The example everybody knows is the first declension with the word-endings for the nominative and ablative which look the same but are different. On the other hand, some forms of different lemmata may coincide. For instance, *oris* is both a form of *ora, ae* and a form of *os, oris*. It can be useful to apply the usual technics of disambiguation to propose the most probable analysis first. Obviously, one has also the perfect homographs, as the two *populus* or the two *levis*, that share the same inflected forms and are completely undistinguishable.

3.1 Statistics on lemmatized texts

Methods based on "hidden Markov models", commonly known as probabilistic taggers, are widely used for disambiguation of the modern languages[42]. They associate a tag to each form that reflect its nature or its function. The Part-of-Speech (PoS) is often used as a tag, sometimes complemented with some other pieces of information. The method relies on the hypothesis that the sequences of tags are characteristic of the language and do not depend on the text, whatever the subject is and whoever the author. Knowing the frequencies of the pairs (form, tag) and the frequencies of the sequences of three tags (second order Markov process), one can compute the probabilities associated with each of the possible sequences of tags for the sentence. Then one assumes that the most probable sequence is the good one, at least the best one[43]. Very high accuracies are obtained with modern languages, where the order of the words in the sentence is rather fixed. It is not demonstrated that the same fidelity can be reached with Latin, where the order of the words is free, or at least much freer than in modern languages.

To start with, one has to choose the tag-set and to do some statistics on a training corpus[44]. A trade-off has to be made for the tag-set. If the tag-set is too small, its disambiguation capabilities will be restricted: for instance, if we just consider the PoS, we will not be able to distinguish the two *oris*, which are both nouns. On the other hand, if the tag-set is too large, the statistics on a finite corpus will be poor. As a training corpus, we got the texts lemmatized and analyzed by the L.A.S.L.A.[45]. They count slightly less than two millions words, each form being associated with a lemma and a code that gives the full analysis[46]. This code cannot be used as a tag, because it would lead to an excessively large tag-set with more than 3,000 different tags. We cut in these codes some redundant information: for instance, for verbs, the group of the conjugation is associated to the lemma and the different persons have different word-endings. We choose to restrict the tag to the PoS associated with the mood for verbs and with the case and number for the declined forms[47]. For each triplet (form, lemma, tag), we counted the number of occurrences in the corpus. We obtained a file with about 150,000 entries. And we did the same for the sequences of three tags, obtaining a file

---

42 See for instance "*A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*" by L.R. Rabiner, in Proceedings of the IEEE, 77, 257-285 (1989).

43 For a more detailed description of a tagger, see "*Probabilistic Part-of-Speech Tagging Using Decision Trees*" by H. Schmid, in International Conference on New Methods in Language Processing, Manchester, UK, 1994 (pp. 44-49). http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

44 It is not a training corpus in the sense used today in Neural Networks and AI. It is a fully annotated corpus on which statistics are performed in a perfectly mastered way.

45 We thank Dominique Longrée and Gérald Purnelle who gave us these texts, the list of which can be found at: http://web.philo.ulg.ac.be/lasla/textes-latins-traites.

46 The gender is absent in the corpus we have treated.

47 The number is needed only to distinguish some forms, mainly in the fourth declension and could be omitted. A lot of tests should be done to optimize the tagset, which are not done for the moment.

with 235,000 entries. These numbers are the primary information sources for the implementation of a probabilistic tagger.

3.2 Double lemmatisation

With the statistical data extracted from the texts lemmatized by the L.A.S.L.A., we have developed a lemmatizer-tagger. The first version of the program began with a sequential lemmatisation. It first looked if the form was found in the file containing all the forms of the texts lemmatized at the L.A.S.L.A. (form lemmatizer). If a word was not found in the file, the code sent a request to Collatinus which was supposed to run in the background on the same computer. Collatinus answered with the possible lemmatizations of this form. If Collatinus was not able to answer (either because it was not running or because it did not recognize the form), then the program asked the philologist who was supposed to supervise the process and waited for an answer.

However, it has been found that this sequential and conditional lemmatisation induces errors and we turn to parallel lemmatisation[48]: the lemmatisation is always done both by Collatinus and by a form-lemmatizer based on the data of the L.A.S.L.A. At the origin of the errors was the fact that as soon as one solution was given by the form-lemmatizer, the program assumed that all the solutions were given. But consider, for instance, nouns where dative plural and ablative plural have the same form. It occurs frequently that for some lemmata only one of these two cases has been found in the L.A.S.L.A. texts. As a consequence, the program assumed that only one tag could be associated with this form, reducing erroneously the tag-sequences to be tried. Then the error propagates due to the mechanism of the probabilistic tagger, forcing the philologist to correct several analyses in the sentence.

The double lemmatisation brings extra work to match, if possible, the lemmata used by the L.A.S.L.A. with those of Collatinus and to remove the duplicates. The correspondence between the two lexica is rather delicate. Just to give some examples, the L.A.S.L.A. distinguishes the two *et*, conjunction or adverb, while Collatinus has a single lemma *et*, with two possible PoS. On the other hand, Collatinus considers (up to now[49]) that *poplus* is a lemma, while the L.A.S.L.A. considers it as a contracted form of *POPVLVS_1*[50]. The correspondence has been established by asking to Collatinus the lemmatisation of the list of forms found in the L.A.S.L.A. files (as mentioned above, the form is associated with a lemma and a code giving the PoS and the analysis). The PoS and the analysis given by Collatinus were compared with the L.A.S.L.A. code. In the best case, the match is unique and perfect, and then the two lemmata are linked. Otherwise, a list of suitors is established and an algorithm tries to sort it out. At the end, a manual check has to be done.

As mentioned above, Collatinus does not split the enclitics *-que* or *-ne* if the word is recognized as a whole. So this possibility has been added in the editor of the annotated text. On the other hand, Collatinus does not search for compound verbal forms, so *amata est* will remain a participle followed by a verb, just as *fortis est* is an adjective followed by a verb. However, in the double lemmatisation, if the compound form has been seen in the L.A.S.L.A. corpus (which is the

48 Independently, Patrick Burns developed concurrent lemmatisation (see elsewhere in this volume).

49 In the last version of Collatinus, we have introduced the possibility to give several forms for a lemma, but we have not yet review all the lexicon to group those forms.

50 As a matter of fact, the lemmata in the lexicon of the L.A.S.L.A. are given in uppercase, with a desambiguisation index if necessary. By convention, proper names and the associated adjectives have always an index, N and A (sometimes O, if there are homonyms as *Pallas, adis, f.* and *Pallas, antis, m.*). Otherwise, the index is present only when there are homonyms and is an integer. In Collatinus' lexicon, the lemmata are written as usual: in lowercase, with an index if there are homonyms (for historical reasons, the index 1 is generally omitted -which is probably not a good idea) and with a capitalized first letter for proper nouns and adjectives.

case of *amata* <*est*>) then the program will propose this solution as the preferred one. This particularity may lead to apparent incoherencies as, for instance, *est amatus* will be recognized as a compound verbal form while *amatus est* will not. But the philologist will have the possibility to add any compound form.

3.3 Disambiguation

The results are sorted according to the frequency, and a first attempt for the lemmatization of the text is obtained by putting together the most frequent individual lemmatizations. This first attempt considers the forms as isolated, independent of their neighbours, and its error rate is expected to be about 20%[51]. Then, the tagger enters to play to take into account the context with a simple statistical model. We have made very few trials: the obtained accuracy was about 88% (exact result, i.e. correct lemma and analysis) and the lemma is the good one in 96% of the cases. As a last step, the philologist can check all the lemmatizations and, if needed, correct them.

It is interesting to note that, though the "context" is described by the sequences of three tags, the choice of the best tags is done only at the end of the sentence or of the text. In principle, all the possible sequences of tags are considered, but a lot of them are skipped[52]. In any case, the choice of a tag can influence the analysis of another word further than two words apart. Conversely, it is important to know how far a "wrong" analysis would spread its consequence. An examination of the list of words shows that slightly less than 40% of the forms are associated to a unique analysis (thus a single tag). Thus, the probability to find two such forms consecutively is 15%, which means that such a pair should be found, on average, every 6 or 7 words. Such a pair splits the text because these unique tags are present in all the tag-sequences, forming fixed points. The fact that we use a second order Markov model implies that the tags that come after a fixed point do not depend on the tags before. Therefore, if the tagger gives the wrong tag to a word, this error will affect some of the following words, but not many. Roughly speaking, it can affect seven words, on average. Obviously, it may happen that a longer series of words can be found between the fixing pairs.

One can think of a "multiplex disambiguation" with another method, which would allow for cross-checking the results. A huge benefit can be achieved if the methods differ sufficiently, even if they are trained on the same corpus. Neural networks and AI are presently very promising in this direction. However, their outputs should be cleaned from the absurdities they can contain. For instance, it has been seen[53] that the output of a neural network program contains "Cum ; cvm ; NOM2 ; Case=Acc|Numb=Sing": the form "Cum" is analyzed as the accusative singular of a noun (lemma) "*cvm*" following the second declension. Clearly, some constraints have to be added to the program. One of the problems with AI methods (in general, this is not specific of this case) is that nobody knows why the program chose one solution instead of another one. This is not the case with HMM where the reason for the choice is always that a probability is larger than another one. By looking closer to these probabilities, it should be possible to associate a "confidence level" to any result. If the larger probability differs from the second one by a small amount, then the confidence level is poor and the philologist should check the result twice. But this is still to be done, and it raises fundamental questions. For instance, what is a small difference in

[51] This figure is evaluated on the training corpus. If we consider the most frequent lemmatization of each form and sum up the corresponding numbers of occurrences, we obtain about 80% of the total number of lemmatized forms.

[52] For details about the pruning method, see the article of H. Schmid quoted in note 43.

[53] We shall not mention where.

probabilities? How can the program, which does not understand what it is reading, know where the difficulties are?

From a more theoretical point of view, it would be interesting to study the sequences of tags to search for correlations. If the order of the words were completely free, one would expect no correlation at all and the tagger would give the same result as a frequency-based lemmatizer. The correlations and the efficiency of the tagger are linked, and the study of the former will give information on the limits in the accuracy. As for the previous point, this work is still to be done. And both points may well be correlated.

3.4 Comparison

The content of this section is mainly subjective and speculative. As a matter of fact, nobody will ever lemmatize the same text with each of the two proposed tools. It would mean to do twice the job with no benefit.

The traditional procedure for preparing L.A.S.L.A. files is semi-automatic: the lemmatizer proposes to the philologist all the analyses known by the L.A.S.L.A. dictionary for each of the forms in the text. The philologist selects the right analysis, or, eventually inserts manually the correct analysis. The analyses are proposed in an order depending only on the morphosyntactic code, and not on their frequency or on their likeliness in that context.

On the contrary, the tagger proposes the most probable analysis, and therefore the role of the philologist is essentially to correct the results of the analysis proposed by the tagger. This accelerates the work, but also changes the kind of human mistakes that can be done. On the one side, the traditional L.A.S.L.A. procedure induces human mistakes caused by the similarity of the possible morphosyntactical analyses, represented by similar alphanumerical codes. The philologist may mistake an accusative for a nominative, or an ablative for a dative, or pick the wrong mood or tense for a verb. It is highly unlikely that, in case of homographic forms, like for instance *salis* (2nd person of the present indicative of *salio*, or genitive from *sal*), the user would select the verbal analysis instead of the nominal or vice versa. On the other side, the tagger may be lead to such an erroneous choice, but the mistake shall remain unseen by the philologist. Indeed, since the philologists expects, for instance, a genitive, he may think that the form is unambiguous, because the possible analysis as the indicative of the verb *salio* may not occur to him in that context. Therefore, the attention may drop, and the tagger's mistake may be left unseen. With the traditional method, the user would hardly mistake the analysis of the verbal form with the one of a substantive. When using the tagger, on the contrary, the philologist is more conscious of the necessity of checking the proposed solution for clearly potentially ambiguous forms, such as datives/ablatives, and will thus probably pay high attention to the correction. At the moment it is not possible to verify which of the methods causes more human mistakes, therefore it is not possible to draw any conclusion on this topic. The two methods are synthetically compared in Table 1:

L.A.S.L.A. Encoding Initiative	*Collatinus-L.A.S.L.A. tagger*
Preparation of the text	
The text is prepared by an operator from the L.A.S.L.A.	The text is loaded directly in the program, with a minimal standardization in the splitting in lines/ paragraphs/chapters...
Pros: Initial control of the edition, of the splitting etc...	**Pros**: The philologist can start to work immediately. He/she has the possibility to correct/change the references and the text during the lemmatisation.
Cons: Possible delays, independent of the will of the philologist.	**Cons**: Possible use of texts (for instance, available on internet) without any indication of the reference to the edition.
Comment: The tagger offers more flexibility, but requires more care and knowledge about the mechanisms of reference and the choice of the edition.	
Choice of the analyses	
Proposition of all the known analyses, without any priority.	Proposition by default of the “best” solution, together with all the other possible analyses.
Pros: The philologist has to read carefully all the given analyses to select one of them.	**Pros**: Fast processing and several cases are solved automatically.
Cons: Constant concentration (even for the simple cases). Slower treatment.	**Cons**: The default choice may be wrong and still escape the philologist's attention.
Comment: An evaluation of the error rates achieved with the two methods has to be done. It is a difficult task from the methodological point of view because it is not the philologist who is evaluated, neither the complexity of the considered text.	
Dictionary	
The dictionary is based on the Forcellini. The addition of new lemmata is controlled by the PI at the L.A.S.L.A.	The dictionary is based on Gaffiot and Lewis & Short. A personal lexicon is added.
Pros: Internal coherence for the whole corpus of the L.A.S.L.A. and also in the propositions given in the program.	**Pros**: More extended lexical base. New entries can be added simply. Distinction between lemmata known by the L.A.S.L.A. (in uppercase) and those from Collatinus (in lowercase).
Cons: Frustration of the manual insertion of new lemmata/analyses. Risk of error in the repetition of this task.	**Cons**: Risk of incoherence with the L.A.S.L.A.'s corpus. Possibilities of unseen doublets or errors in the indices.
Comment: Strong advantage in the speed of the tagger. If the personal dictionaries were checked and inserted in the L.A.S.L.A. dictionary, it would increase its size rapidly.	
Final treatment	
Usually, the treated text is checked (often by another philologist). Correction of the printed index and insertion of them by an operator. Production of the final file, by an operator, at the end of the process (for instance, several books).	The generation and the correction of the index are left to the philologist. The output file is immediately in the standard APN format which makes it usable at once.

L.A.S.L.A. Encoding Initiative	*Collatinus-L.A.S.L.A. tagger*
Pros: Rigorous verification, in part on printed material.	**Pros**: The file can be studied as soon as it is completed, without having to wait for the completion of the entire work (if formed of several books).
Cons: Possible delays in the processing (in part independent of the philologist's will).	**Cons**: Risk of a less careful verification.
Comment: Working with the tagger appears to be a more personal work, with more responsibilities but more independence and flexibility.	
Conclusion: For a work to be completed in a finite amount of time (e.g. for a PhD thesis), the speed of the tagger is a key element. The philologist at work has a complete controle of all the steps, but also (as a consequence) a larger responsibility. On the long time scale, the traditional method is safer for the coherence of the L.A.S.L.A. corpus. However, nothing impedes an extra checking of the output of the tagger (by a second philologist) to ensure its quality. The coupling of the two methods could lead to a significant increase of the L.A.S.L.A. corpus and dictionary.	

Table 1: Summary of the differences between the two NLP tools.

4 Conclusion

In this article, we have presented part of the work going on at the L.A.S.L.A. and in the Collatinus' developing group. We have also put some emphasis on their collaboration and compared the two approaches for the lemmatisation and analysis of new Latin texts. We underline the pros and cons of each of them. A kind of trade-off has to be found between speed and precision.

However, the required precision or the tolerable error rate may depend on the envisioned application and remain an open question. Obviously, a perfect lemmatisation, with no error at all, is desirable, but probably not needed. Most of the applications are of statistical nature, which means that they contain an intrinsic degree of uncertainty which can often be determined with error-bars, but seldom given or understood. In this context, what is (or would be) the consequences of a few remaining errors? It is difficult to evaluate, but even more difficult to measure. Due to the lack of realistic objectives (with upper limits on the acceptable error rate, for instance), we stick to perfection.

Authors information:

Philippe Verkerk[54], Laboratoire de Physique des Lasers, Atomes et Molécules; Bâtiment P5; Université de Lille; F59655 Villeneuve d'Ascq CEDEX; FRANCE. Philippe.Verkerk@univ-lille.fr

Yves Ouvrard, retired professor of “Éducation Nationale”. Yves.Ouvrard@collatinus.org

Margherita Fantoli[55], Laboratoire d'Analyse Statistique des Langues Anciennes; Université de Liège; Place du 20 Août, 7; B-4000 Liège; BELGIQUE. mfantoli@uliege.be

Dominique Longrée, Laboratoire d'Analyse Statistique des Langues Anciennes; Université de Liège; Place du 20 Août, 7; B-4000 Liège; BELGIQUE. dominique.longree@uliege.be

[54] Also « Collaborateur de l'Université de Liège ».

[55] Present address: Amsterdam